The Role of Natural Language in Accident Investigation and Reporting Guidelines

Kimberly S. Hanks, John C. Knight
Dept. of Computer Science, University of Virginia
151 Engineer's Way, Charlottesville, VA 22904-4740, USA
{ksh4q | knight}@cs.virginia.edu

C. Michael Holloway
NASA Langley Research Center
MS 130 / 100 NASA Road, Hampton, VA 23681-2199, USA
c.m.holloway@larc.nasa.gov

Abstract: The need to learn from incidents and accidents resulting from software failure to improve the development process and reduce the incidence of such events mandates a rigorous discipline of forensic software engineering. The proliferation of assumption in the notions and representations of critical concepts during a software process is a barrier to developing this discipline. This is true not only of documents such as requirements statements and investigation reports, but also of the guidelines that dictate how investigation of failures should take place. The goal of investigation guidelines is the production of a report with certain properties, and proliferation of assumptions in the statement of such guidelines impairs the attainment of this goal. Drawing on linguistics and cognitive psychology, we earlier motivated an approach to improving the natural language of requirements statements. In this paper, we examine the issues surrounding the natural language in which investigation and reporting guidelines are written, and suggest ways that they can be demonstrably and systematically improved with our approach. The issues and approach are demonstrated using the NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping. Deficiencies are explored, potential consequences discussed, and a strategy for systematic improvement of the document is outlined.

Keywords: natural language, investigation guidelines, forensic software engineering.

Introduction

Incidents and accidents that can be attributed to software failure often result in tragedies and other losses. The need to learn from these events grows more critical as software systems become more complex and the ways they can fail become less intuitive. This need mandates the development of a more rigorous and systemic approach to forensic software engineering.

In exploring the issues surrounding the development of a systemic approach to forensic software engineering, Johnson, 2000 recognized that the proliferation of assumption and other communicative problems throughout the software process were recurring themes in investigations of accidents. Importantly, he also demonstrated that these same deficiencies that plagued the process also plagued the reports and recommendations resulting from these investigations. These deficiencies raise two issues. First, the potential for assumption and other miscommunication during the development process, especially in the early stages of a software project, can allow invalid conceptions of elements of the system to enter and persist, possibly leading to failures. Hayhurst and Holloway, 2001, among others, argued that requirements is a communication problem, and thus that poor requirements result from poor communication, and Lutz, 1993 showed that the majority of safety-critical errors in the systems she examined were introduced at the requirements stage. Further, unstated or unclear motivations for requirements decisions impair the ability to analyze causes. Second, the potential for assumption and other miscommunication in the reports and recommendations resulting from investigations of such failures can render such documents of little use. For example, if the analyses contained in a report are based on misconception, then a valid analysis has escaped recognition, and if the recommendations suggest ideals that are assumed to be achievable but are in reality impossible (as documented in Johnson, 2000), then time and energy that could be applied to exploring new avenues for progress is likely to be wasted in the service of unattainable perfection.

To these we add a third issue: there exists the same potential for assumption and miscommunication in the statements of guidelines that prescribe the activities and artifacts associated with incident and accident investigation and reporting. In contrast with the requirements for a software system or the reports resulting from investigations, guidelines represent meta-statements; they define the form and content of a class of instances, whereas requirements and reports are instances of classes. The purpose of these meta-statements is in large part to standardize the results of investigations, such that, as a data set, the results can be compared with one another and analyzed for trends. In other words, the intended value of guidelines is that they predictably generate artifacts with properties that are useful to forensic software engineering. The potential for assumption and miscommunication in such guidelines impairs the likelihood that, for example, two different investigation teams will come to substantially the same conclusions while following the prescribed process, that is, this potential impairs predictability and the value of resulting documents as a data set. This creates an additional area of focus within a systemic view of forensic software engineering, in which reduction of the potential for assumption and miscommunication in investigation and reporting guidelines is a necessary task if the big picture and the role of communication throughout it are to be improved.

In previous work, we examined how the ways that humans innately use natural language render statements of requirements incomplete, inconsistent, and open to misinterpretation (Hanks, Knight, and Strunk, 2001). This analysis exploited results from cognitive linguistics that detail the ways in which humans organize and communicate conceptual information. We extended this model to account for the breakdown that occurs in communication of information across boundaries of domain expertise, breakdown that is implicated as a major limiting factor of the quality of large and complex software systems (Curtis, Krasner and Iscoe, 1988).

Miscommunicated requirements, as noted above, are themselves a detriment to forensic software engineering (Johnson, 2000). However, the model by which we analyze and characterize communicative breakdown in requirements can also be applied to investigation and reporting guidelines, as well as to the reports and other documents that are generated. Implications of the model suggest ways to improve the use of natural language in all of these areas. In this paper, we treat particularly the problem of incomplete, inconsistent, and ambiguous guidelines for the investigation and reporting of incidents and accidents. We begin with a more detailed discussion of the issues particular to guidelines and their communication. Next, we review the analysis model, which is followed by an overview of the approach that the model motivates. We then provide a case study of the language used in the NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping. Finally, we describe a plan for improvement of the document through systematic reduction of incompleteness, inconsistency, and ambiguity.

The Role of Natural Language in Investigation and Reporting Guidelines

It is important that the investigation of any incident or accident be effective and efficient. A variety of techniques and procedures has been developed to assist with these goals, and many organizations have developed guidelines for investigation and reporting. An important objective of guidelines is to facilitate the creation of results that have *predictable* properties, in particular, ones that facilitate comparison among results of multiple investigations in order to observe patterns. Guidelines help to ensure that results are comparable by prescribing processes, procedures, and formats. Figure 1 illustrates the relationship of guidelines to a number of activities surrounding and directed by them. The structure that emerges when one considers the role of guidelines is that they are meta-documents—they are used to instantiate particular investigations and reports. Any deficiency in the guidelines, even one that seems unimportant, could have an extensive negative effect if it leads to significant imperfections in many investigative or reporting instances.

Deficiencies in guidelines do not need to take the form of factual errors to have a substantial effect. Ambiguous statements in guidelines can be extremely serious because the multiple meanings lead to results that differ from one investigation to another, thereby precluding the goal of predictability. Further, statements that are incomplete affect predictability because the incompleteness leads to instantiations that are either themselves incomplete or completed in an ad hoc manner. Finally, inconsistency in guidelines can result in instantiations that differ because of different interpretations arising from the inconsistency during instantiation.

As we show in the next section, natural language and complex cognitive structures serve our everyday needs as humans in a way that is not consistent with the goals of precise guideline statements. Unless this issue is addressed, the many opportunities for misunderstanding that are inherent in our unrestricted use of natural language can have disastrous effects in situations where completeness, consistency and precision are essential.

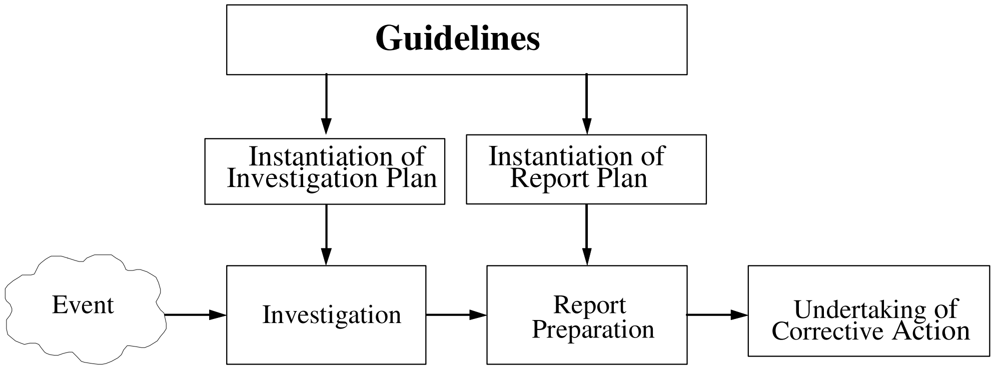

Figure 1: The role of guidelines.

The Category Model and Its Implications

Research in linguistics and cognitive psychology has demonstrated that the universe of semantics understood by any person is organized into a collection of structured entities called *cognitive categories* that possess various properties (Rosch and Lloyd, 1978, Mervis and Rosch, 1981, Ungerer and Schmid, 1996, and Langacker, 1990). For our purposes, these cognitive categories can be defined as follows:

> *Cognitive categories are collections of mental representations of entities encountered or imagined by an individual that are judged by that individual to be sufficiently similar to each other to count in some partitioning of reality as being the same.*

An individual's categories are formed as a result of his or her accumulated experience. Since there are many possible partitionings of reality that are useful to us in our interaction with the world, any entity can be a member of more than one category—which category depends on the factors considered to be significant for the task or experience at hand.

Categories are collections with internal structure based on a notion of resemblance or similarity to a prototype characterizing the category (Figure 2).

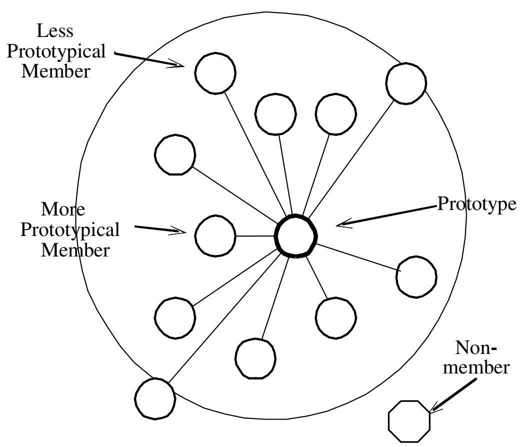

Figure 2: Structure of a cognitive category.

Members of a category that are closely clustered around the prototype bear stronger resemblances to it and instances further away have less resemblance. For example, a helicopter is a member of the *aircraft* category but is less aircraft-like than, for example, a Boeing 747. However, by being a member of a given category, regardless of how prototypical, an instance is associated with a collection of attributes common to members of that category. By communicating a single term, such as helicopter, a speaker conveys many attributes to a listener. The listener does not need additional communication to know that the entity to which reference is made has a pilot, an engine, and flies. This aspect of human communication is known as *cognitive economy*.

Human communication in routine circumstances would be almost impossible without cognitive economy. The bandwidth of human communication is quite low, but because of cognitive economy a great deal can be communicated using that low bandwidth. Without it, we would be forced to describe all of the details about every entity we ever wished to mention. Clearly this would be impractical.

The efficacy of cognitive economy rests on shared experience. Using the word helicopter when speaking to someone who has no experience with helicopters is useless. Provided those engaged in communication (whether written or oral) have similar experience, all is well. If their experience of the domain at hand differs, even if the difference is slight, then the attributes associated with a term will not be the same. In routine communication this is not a major issue, but when using natural language in a context that requires extreme accuracy and precision, cognitive economy is the source of many invalid assumptions.

The issues that arise are of two types. In the first, the listener recognizes that, although a term used is familiar, some aspect of the communication is not clear. In this case, the listener can investigate the meaning of the problematic term by asking questions and thereby negotiate a closer match between his own conception and the speaker's. The second type of flaw is far more serious, and it occurs when the listener, through lack of or inattention to cues, does not recognize that there is a problem. In this case, the listener *assumes* that his understanding of a term and all the associated attributes implied by cognitive economy are the same as those of the speaker when they are not. All subsequent activities and behaviors will therefore be in the context of this misunderstanding.

In addition to having internal structure, the categories we possess are collectively organized into a hierarchy of specificity, with more general categories at higher levels and very constrained categories at lower levels. This hierarchy is one of inclusion, meaning that many low-level, highly-constrained categories are collected under the umbrella of a more general category, several of these more general categories are collected into still more general categories, and so on (Figure 3).

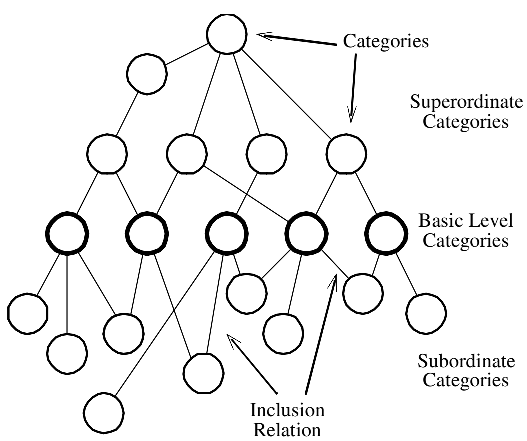

Figure 3: Hierarchical structure.

There is evidence to suggest that a particular level of this hierarchy has a special importance in our perception of the world (Rosch, Mervis and Gray, et al., 1976, Rosch and Lloyd, 1978, and Ungerer and Schmid, 1996). This level is intermediate; it is neither a very general way to describe a thing nor very specific (for example, *airplane* rather than *vehicle* or *Boeing passenger transport*), and its special role is evidenced by features particular to it in our acquisition and use of categories. It has been termed the *basic level* and represents the categories first acquired by children, the categories first evoked when classifying a newly encountered entity, and the categories used when introducing a new category into communication.

Analysis of empirical data has demonstrated a property of basic-level categories that supports their importance in communication. The basic level is that level of the hierarchy at which elements of any given category share the most features with each other and the fewest with members of other categories (Rosch, Mervis and Gray, et. al, 1976, Rosch and Lloyd, 1978, and Ungerer and Schmid, 1996). Unfortunately, this property further complicates the reliable communication of domain knowledge. This added complication derives from the nature of the domain-specific categories used by experts. Domain experts have accumulated experience that results in their associating more attributes with domain-specific categories than a non-expert would. This is an obvious result of the very fact that an expert is an expert. The additional attributes provide more dimensions along which to collect and differentiate entities resulting in certain of these categories being basic in the expert's category hierarchy.

The implication is that experts tend to see what are commonly lower-level, more constrained categories as basic in their own hierarchies, and to use them in ways that basic-level categories are used. On being presented with a new entity in his domain, an expert is likely to associate it with a more constrained category than would a non-expert. Similarly, on using a domain-related entity in communication, the expert is likely also to invoke a more constrained category. This means that in addition to experts and non-experts possessing more and less constrained versions of certain categories, the denser expert versions are more likely to come up in discussions in a specific context because, to the expert, they are at the basic level. This results in more misalignment between the categories used by experts and non-experts than would occur because of the backfiring of cognitive economy alone.

To review, the mechanics of linguistic breakdown in the communication of domain knowledge can be characterized as follows. First, the benefits of cognitive economy that allow us to communicate adequately though somewhat imperfectly in our routine activities lead not only to the potential, but the *likelihood* that erroneous assumptions will be made in high precision, technical communications. Second, added complications arise from the specific categories that domain experts regard as basic because they are at a different (lower) level than the categories regarded as basic by the non-expert. Communication across a domain boundary, communication that is essential if investigation guidelines are to be sufficiently comprehended by investigators with diverse expertise, embodies exactly the properties that cause our natural machinery to fail. It is not a part of human nature to get this kind of communication right without serious and explicit intervention.

Approach

Using insights gained from a linguistic analysis of breakdown in domain knowledge communication, we developed an artifact designed to manage and contain the potential for such breakdown. Consider the case in which any two people with differing levels of expertise with regard to a topic are communicating regarding an entity relevant to that topic. Further assume that one or the other has in fact recognized that a breakdown is occurring. This is not representative of the more dangerous situation of no breakdown being signaled, but motivates a strategy for preventing the breakdown from occurring in the first place. In the case where breakdown is recognized as it is happening, the usual course of action taken by the interlocutors is to execute clarification activities. These activities generally take the form of paraphrasing the offending term with another term or phrase for which the sender believes the receiver is likely to possess a more compatible category topology or topologies. If this paraphrase contains terms that also invoke misaligned categories, these terms can further be paraphrased, and so on. Comprehension is recursive; we comprehend a new idea when we can put it in terms of other ideas that we already comprehend. This insight provides a direction for dealing with the problem produced by reliance on assumption in communicating domain knowledge.

Our approach is to introduce a highly structured mechanism, called the *domain map*, into communication activity that requires accuracy and precision. The map stores definitions of domain-specific terms, and documents their recursive dependence on definitions of other terms for their comprehension. It is intended to provide a systematic and complete repository of relevant domain semantics built according to the principle of making the implicit explicit. Further, once constructed, it is to serve as the exclusive point of reference for such semantics where the content of the document is concerned, providing a consistent picture to its users, for example, members of an investigation board.

Specifically, the domain map is to be constructed using a starting point of some recorded natural language, for example, a written document such as an early draft of guidelines. This body provides a corpus representing an instance of the language used to talk about the domain in question and the constraints to be placed on it. In an iterative process, experts in the various areas that contribute to the guidelines, and representatives of non-experts who are likely to be users, cooperate to partition this corpus into terms identified respectively as *domain*, those that have domain-specific meaning, and *common*, those that are unlikely to invoke relevant differing assumptions between experts and non-experts. One focus of our parallel work is refining this partitioning activity to be based rigorously on specific membership criteria for these sets. However, early experiments have been quite successful even with partitioning accomplished in an ad hoc, intuitive manner (Hanks and Knight 2002).

Once the initial *domain* set is constructed, each of its elements can be defined precisely, again in an iterative process executed by cooperating experts and non-experts. For example, a non-expert might make a first attempt, which the expert would then examine and revise. A stipulation placed on the process is that, for each term, the parties must agree that they have converged on the same understanding, as interpreted from the definition, before moving forward. This implies that both parties must have the same understanding of *each term* in the definition. Thus, terms upon which the initial term depends must themselves be classified as *domain* or *common* and defined as necessary. This realizes the recursive nature of comprehension, and forces the parties to trace these dependencies.

The bottom of the recursion is defined by design, and thus the trees representing term definitions bottom out with terms that are accepted without definition. This is the purpose of the *common* set; by virtue of its construction, it consists of those terms deemed to represent knowledge common to those both inside and outside the domain, and its use is to provide the source lexicon on which all domain definitions must eventually depend. Thus another stipulation is necessary: no cycles are allowed in the chain of dependencies associated with any term. A cycle would indicate that the recursion would never terminate, i.e., that a common understanding would not be reachable. Parties must therefore negotiate their removal, thus addressing circularities that might not otherwise have been recognized.

Our linguistic model and approach have been shown to have value in analyzing linguistic deficiencies and improving the quality of software requirements statements (Hanks and Knight 2002). Insofar as investigation and reporting guidelines can be seen as another kind of requirements statement, we extended the application of our model and approach to this area. We next discuss this extension.

Case Study: NASA Procedures and Guidelines

Linguistic Deficiencies: To illustrate the ways in which the use of natural language encourages the proliferation of assumption in investigation and reporting guidelines, we have conducted an analysis of the glossary section of the current version of the NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping (hereafter referred to as "NPG"). We chose a subsection of the document for practical purposes of illustration; the entire document is 176 pages and a complete analysis is beyond the scope of this work. However, we found the glossary particularly compelling, since this is the section in which the document authors were *specifically tasked with* communicating explicitly the meanings of critical terms. Our analysis indicated a number of deficiencies that impair the value of the document to those attempting to realize its intended purpose.

First, a partitioning pass over the glossary text indicated incompleteness in the set of terms chosen to be defined. Since we were concerned for the moment with the 6 pages of glossary alone, this incompleteness does not yet consider problematic terms in the 170 pages of additional text; it refers rather to the necessity of access, either through prior knowledge or definition, to the meaning of terms upon which the understanding of glossary entries relies. Since this section is a glossary, we should reasonably expect that upon reading a given definition, the meaning of the defined term should be clear. If the definition includes other terms that might be problematic, we should expect to find definitions for these as well. However, *every* definition included in the original glossary contains terms that potentially have several meanings, but which are not themselves included in the glossary. For example, a term that is invoked in many of the definitions is *investigation*, indeed, it represents a concept central to the purpose of the document, yet it is not defined in the glossary. While it is true that speakers of English might collectively have a general idea of what an investigation might entail in the abstract, in this context it refers to a specific collection of activities, and a brief enumeration and description of these activities would render understanding of the terms that rely on this use of *investigation* that much more meaningful to users of the document. Further, such an enumeration would discourage users from assuming the inclusion or exclusion of activities that characterize other investigations with which they are familiar, an assumption motivated by cognitive economy, but one that might be invalid.

Second, further incompleteness was represented by occurrence in passages throughout the remainder of the document of additional terms in need of definition. We can thus say that the glossary is both locally incomplete, i.e., that the definitions included are not themselves completely grounded, as well as globally incomplete, i.e., the set of terms chosen from the main text to be represented in the glossary is not comprehensive. For example, *proximate cause* occurs in the main text, and though it may have a standard definition among those with expertise in causality analysis, the individuals tasked with using the guidelines do not all possess this expertise. Further, the authors *did* explicitly include terms like *dominant root cause* and *significant observation*, among others, indicating that they saw a need to distinguish such concepts from one another. *Proximate cause*, however, was overlooked.

Third, in addition to incompleteness, there are several forms of inconsistency in the document. An example within the glossary surrounds the definition of a NASA Mishap, which includes an enumeration of possible types, corresponding to severity estimations: "Type A Mishaps, Type B Mishaps, Type C Mishaps, Mission Failures, or Incidents". A reasonable interpretation of this list might be that the set of all NASA Mishaps can be partitioned into five mutually exclusive subsets. However, upon reading the definition for *Mission Failure*, we find that it refers to "[a] mishap of whatever intrinsic severity" that also possesses certain other properties. This directly contradicts the understood mutually exclusive partitioning; a *Mission Failure* can apparently also constitute, for example, a *Type B Mishap*. It is not clear how to resolve the contradiction, and if a user happens to refer to only one of these definitions, he would likely not even recognize that there *is* a contradiction.

Inconsistencies relating the glossary to the remainder of the document are present as well. A number of concepts invoked in the glossary appear to be represented by different terms, or sets of terms, in different locations. For example, *NASA Mishap* has an explicit entry in the glossary, but throughout the text of definitions as well as in the text of the complete document, simply *mishap* is invoked. Since mishap has a common lay usage, a user might reasonably read it as such. A more vigilant user might suspect a domain-, i.e., NASA-specific definition, but upon looking up *mishap*, would find no glossary entry and thus also reasonably assume common usage. Only had he looked up *NASA Mishap* would he have located the presumably intended meaning, but how is he to know to look there? A similar example involves the occurrence in both the glossary text and main text of *injury/illness*, coupled with the explicit glossary entry *Lost-time Injury/Illness*. The lookup problem is the same; a user wondering what constitutes an injury or illness will not find an entry for *injury/illness* in the glossary and might assume criteria based on other experience. However in this case, it is even less clear from surrounding context whether these representations do in fact refer to the same concept, that is, are there injuries/illnesses that do not cause lost time? Such inconsistencies of representation (which in the latter case may be masking incompleteness) hinder the value of the document for directing the analysis of events by making it more difficult to classify and relate objects in the world that are of interest in an investigation.

Fourth, the glossary text includes numerous terms that have abstract common meanings for which most speakers of English have similar notions. However, the abstract meanings have little value when placed in a specific context unless criteria are provided that parameterize these meanings within that context. For example, the terms *appropriate*, *authority*, *generally*, *ordinarily*, *significant*, *similarly*, *major*, *minor*, *basic*, and several others all occur one or more times within the glossary text alone. Some of them occur many times, and a number of instances as well as additional such terms were found in the remainder of the document with only a cursory search. A section addressing the composition of NASA Mishap Investigation Boards states that "[m]embers shall have sufficient experience and technical expertise" to uphold their responsibilities, but there is no indication of how *sufficient* or even *expertise* are to be qualified or quantified. These terms are all quite transparent in the abstract sense, but since they are relative descriptions or measures, they require reference points to have any useful meaning in a given environment or domain. This renders these terms in fact domain-specific once they are actually invoked in a context, and they thus require definitions and encourage assumption without them.

So far, we have concerned ourselves primarily with individual terms, however the magnitude of the problem becomes obvious when we try to deal with several terms at once. Presented here is a passage from the document addressing the intended form and content of reports that result from investigations (and recall that we are not quite sure what exactly investigations entail). The passage is followed by a selection of indications of its insufficiency for directing the construction of such a report.

> 3.7.5 The mishap investigation report will contain a description of the structured analysis technique used by the mishap investigation board or investigator for assuring all causative possibilities are explored. The mishap investigation board or investigator will document the what, when, where, and why of the mishap investigation report. The focus and priority of the investigation report is the determination and discussion of the root cause(s) of the mishap. The report will also include significant observations, findings, and recommendations. The report will include proposed corrective actions if requested in the appointment letter, and proposed lessons learned topics for future development. The report should be technically accurate, properly documented, well defined, easily understood, and consistent with the format in Appendix H or as specified by the Appointing Official.

First, neither *mishap investigation report* nor *structured analysis technique* are defined. *Structured analysis technique*, in particular, has a definition specific to software engineering, but which is almost certainly not the meaning intended here. Since an investigation board is likely to include both software engineers and non-software engineers (and these groups understand these terms differently), a board is not likely to begin with a coherent notion of what they are to describe. Further, before developing such a coherent notion, they would first have to *recognize* this inconsistency in their experience of the terms, which they might not do until much effort has been invested under faulty assumptions. An explicit definition could reduce or avert such misappropriations and inefficiencies. In addition, since any definition of *mishap investigation report* is likely to include *structured analysis technique* among the elements such a report must describe, a valid definition for *structured analysis technique* is necessary in order to ground the definition of *mishap investigation report*.

An explicit definition of *structured analysis technique* might also encourage critical reflection on the value and limits of such techniques; note the assumption in the above passage that the use of a structured analysis technique can "assur[e] all causative possibilities are explored." While attaining this assurance might be a useful ideal for motivating and directing analysis activities, it is quite impossible to do so perfectly and demonstrably in the complex environments with which we are concerned. Johnson, 2000 recognizes such statements in the recommendations made in existing accident reports. For example, a report on the deficiencies of the London Ambulance Computer Aided Dispatch system contains the recommendation: "A critical system such as this...must have totally reliable software." Johnson states "It is impossible by any objective measures to achieve total software reliability, contrary to what is suggested..., [and] to suggest that this is possible is to completely misrepresent the state of the art in safety-critical software engineering." We must be wary of similar such assumptions and suggestions in the guidelines we provide to investigators. It is counterproductive to assign exercises in futility and to forego a number of forms of progress in the quest for an unattainable perfection.

In addition to the incompleteness represented by the unavailability of certain definitions, an instance of the "mishap" inconsistency described earlier is also found here. Further, though definitions for *root cause*, *corrective actions*, and *lessons learned* are provided, a prescription for the form and extent of the required description of these is not. This demonstrates further incompleteness.

Also represented are additional instances of abstract common terms in need of contextual parameterization. For example, what constitutes *technical accuracy* in this domain? Similarly for *proper documentation*, *well-definedness*, and *easily understood*. *Easily understood*, in particular, begs the question of audience, i.e., the diversity of readers of these reports and their expertise.

Finally, it is not clear whether Appendix H or the Appointing Official is the final arbiter of format; if both are consulted and they disagree, which is to be attended?

With this much potential for misunderstanding and therefore unpredictability of the result contained within a single paragraph of the NPG, it is clear that there is a linguistic problem to be addressed. Next, we outline a strategy for improving the document by systematically reducing the amount of incompleteness, inconsistency, and ambiguity contained therein.

Strategy for Improving the NPG: In this paper, we have shown that the NPG in its current form has deficiencies. The deficiencies have a basis in the way that humans innately use natural language, and derive from the fact that our cognitive heuristics are optimized for situations in which communicators share experience. Communication across a domain boundary is the pathological case that breaks these heuristics; they work in the common case by exploiting assumption, but they are the source of pervasive and often dangerous miscommunication in cases where shared experience is lacking.

We believe the NPG can be improved through the application of methods originally developed for raising the quality of requirements statements. In particular, a domain map, such as was described earlier, can systematically reduce the amount of incompleteness and inconsistency present in the NPG. The following activities, intended to be undertaken by cooperating domain experts (in this case authors or those capable of authoring the document) and analysts tasked with implementing the improvement project, represent a strategy for this systematic improvement.

We would begin with the existing glossary and its local incompleteness and inconsistency. As recognized above, not only are there many terms from the main text not explicitly defined, but the definitions that *are* provided are not themselves completely grounded. We would first complete the glossary in this down-dependency direction, that is, add to the glossary those problematic terms that are present within the definitions of already-defined terms, for example, *investigation*. It is further necessary in this step to examine added definitions for their own problematic terms. The point is to produce a domain map for the glossary that is as close as possible to being internally complete. To address, next, the inconsistencies in the glossary, requires that all uses of defined terms be checked for usage consistent with the provided definitions. For example, the conflict in the definitions of *Mission Failure* and *NASA Mishap* above must be resolved. In addition, the cases of multiple representations of single concepts must be addressed; if *mishap* and *NASA Mishap* refer to the same concept, their representation is to be standardized, likewise for *injury/illness* and *Lost-time Injury/Illness*. These changes allow a glossary that is much closer to being internally, or locally consistent. Increasing local completeness and consistency approaches the goal of making all entries transparent to any user likely to require use of the glossary. These local steps alone improve on the original by systematically addressing the terms that the authors themselves, using even intuitive methods, believed required definitions.

Once local completeness and consistency have been addressed, a more extensive project is to address global incompleteness and inconsistency, that is, the necessity for the intended meanings of any terms occurring in the main text to be transparent to any user likely to be reading the document. Given the size of the document, this effort is likely to require a non-trivial investment of effort, but since the document has a lifetime of approximately five years (NASA QS/Safety and Risk Management Division, 2000), and further, since much of its content persists through version updates, this investment can be amortized. Further, the return is greater confidence in the value of the document to effectively direct, and standardize the artifacts resulting from, investigation and reporting of incidents and accidents.

To address global completeness, we would partition the entire text into two sets: (1) those terms that have domain-specific meaning not likely to be transparent to all users and thus requiring explicit definition; and (2) those terms for which the common lay usage is what is to be understood. All unique terms in the first set are added to the glossary as independent entries. Definitions are constructed for these terms, and those definitions are then processed as before to maintain local completeness and consistency. To address global consistency, all uses in the main text of defined terms must be checked against the definitions provided. In addition, as before, multiple representations must be standardized.

Much of the process just described can be simplified through automation via support tools. In parallel to formulating theory and conducting analysis, development of such tools is underway at the University of Virginia.

The recognition of note is that this process, undertaken manually or otherwise, is systematic. Every term is processed, the form of processing is motivated by linguistic insight regarding the ubiquity of assumption, and the team tasked with improving the document is directed by the process to give specific attention to all terms that might cause problems through the potential for invalid assumptions about their meanings. Specific constraints direct decisions during processing, for example, that there be no cycles in the dependencies among terms; this forces not just the constructing of necessary definitions, but evaluation of their usefulness. The result is a more comprehensive and considered representation of meaning essential to the effective undertaking of an investigative task.

As with all Agency directives, guidebooks, and handbooks, NASA has procedures in place that must be followed to suggest changes to official documents. If we choose to follow through on the approach to changing the NPG just suggested, we plan to formulate our recommendations for changes in the style required by those procedures.

Summary

The proliferation of assumption in the notions and representations of critical concepts during a software process is a barrier to effective forensic software engineering. This is true not only of the content of reports generated during accident investigations, but also of the requirements and designs describing software to be built, and, as we have examined, the guidelines that dictate at a meta-level how analysis and investigation of failures should proceed. The goal of investigation guidelines is the production of a report with certain properties, and proliferation of assumption in the statement of such guidelines impairs the effective realization of this goal. We argued that the way humans innately use natural language encourages the proliferation of assumption in environments where individuals with differing experience and domains of expertise must communicate. We further argued that our cognitive heuristics are so ingrained that progress in overcoming their challenges will not be made without well-founded and structured intervention. The linguistic model we provide motivates much of the structure that this intervention must take in order to be effective. The result is a systematic approach to reducing the incompleteness and inconsistency of investigation and reporting guidelines, leaving demonstrably less room for assumption, and allowing such guidelines to better serve their purpose. The issues and approach were demonstrated using the NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping. Deficiencies were explored, their potential consequences were discussed, and an approach to systematic improvement of the document was outlined.

References

B. Curtis, H. Krasner and N. Iscoe (1988). A field study of the software design process for large systems. Communications of the ACM (31)11:1268-1287.

K. Hanks and J. Knight (2002). An experiment in applying linguistic insight to improve requirements. University of Virginia Department of Computer Science Technical Report CS-2002-18.

K. Hanks, J. Knight and E. Strunk (2001). Erroneous requirements: a linguistic basis for their occurrence and an approach to their reduction. Proceedings of the 26th Annual IEEE NASA Software Engineering Workshop, 115-119.

K. Hayhurst and C.M. Holloway (2001). Challenges in software aspects of aerospace systems. Proceedings of the 26th Annual IEEE NASA Software Engineering Workshop, 7-13.

C. Johnson (2000). Forensic software engineering. Proceedings of 19th International Conference SAFECOMP 2000, 420-430.

R. Langacker (1990). Concept, Image, and Symbol: The Cognitive Basis of Grammar. Mouton de Gruyter, Berlin.

R. Lutz (1993). Analyzing software requirements errors in safety-critical, embedded systems. Proceedings of the First IEEE International Symposium on Requirements Engineering, 126-133.

C. Mervis and E. Rosch (1981). Categorization of natural objects. Annual Review of Psychology 32:89-115.

NASA QS/Safety and Risk Management Division (2000). NASA Procedures and Guidelines for Mishap Reporting, Investigating, and Recordkeeping (NPG 8621.1).

E. Rosch and B. Lloyd, eds. (1978). Cognition and Categorization. Lawrence Erlbaum Associates, Hillsdale, NJ.

E. Rosch, C. Mervis, W. Gray, D. Johnson and P. Boyes-Braem (1976). Basic objects in natural categories. Cognitive Psychology, 8:382-439.

F. Ungerer and H. Schmid (1996). An Introduction to Cognitive Linguistics. Longman, London.